DOCUMENTATION
INCORPORATED

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

MACHINES AND CLASSIFICATION IN THE ORGANIZATION OF INFORMATION

TECHNICAL REPORT NO. 2

PREPARED UNDER

CONTRACT NO. Nonr-1305(00)

for

THE OFFICE OF NAVAL RESEARCH

DECEMBER
1953

DOCUMENTATION
Incorporated

WASHINGTON 8, D. C.

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

DOCUMENTATION
INCORPORATED

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

ABSTRACT

Classification of information seems to offer a more logical arrangement of meanings and associations than does alphabetical indexing. However, fifty years of library experience with classification systems indicates that classification is useful only within the limited fields of the taxonomic sciences.

The recent rebirth in interest in general classification systems is traced to certain problems relating to the machine storage and retrieval of information; and it is shown that this new interest, together with the new terminology of "abstraction ladders," "semantic factoring" and "categorization," offers little promise of solving the inherent difficulties of hierarchical classification. It is concluded that classification remains a "blind alley" and that other techniques and principles of associating meanings must be found and developed.

MACHINES AND CLASSIFICATION IN THE ORGANIZATION OF INFORMATION

PART I

In previous papers we have recognized that classification, as contrasted with alphabetical subject headings or coordinate indexing, supplies a kind of connectiveness between ideas and provides the possibility of "browsing" through related ideas.

> "Alphabetical arrangement is most convenient for the user who can precisely name the subject of his search in the same terminology as used by the indexer, but may make searching difficult for others. Those users who are not current with the fashions in nomenclature, who are not completely familiar with the subject of search who have only a vague question in mind, etc., can be helped by a system whose arrangement is related to the organization of the field searched."

> "Classification can arrange ideas, not merely words, since meaning can be indicated through position, as well as phraseology where terminology is not fixed. Browsing among related concepts is of course, facilitated by placing them in proximity."[1]

The degree to which any classification system associates ideas is a measure of the effectiveness of

[1] Studies in Coordinate Indexing, pp. 68-69. Documentation Incorporated, Washington, D. C. (1953).

the system. Where classification fails is in the arbitrary disassociations which are imposed on related ideas by the requirements of the system. These arbitrary disassociations are hidden and the searcher who is content with the relationships displayed in the system will obtain only partial information.

In short, for reasons which will be set forth in detail below, classification systems are not truly effective instruments for displaying to the browser or searcher all the ideas in any system which are associated with any given idea with which the system is entered.

This fact was recognized by Dr. Vannevar Bush in the quotation[2] which posed the general problem of the association of ideas as presented in our previous report.

In spite of all the theoretical and practical objections which can be marshalled against classification as a method for organizing information, classification systems are apparently successful in associating ideas torn asunder by alphabetical indexes. This leads many individuals and organizations to ever new attempts to devise classification systems. These attempts have multiplied in recent years, not because of any actual successes on the record, or any new developments in classification theory. They have multiplied because the machine searching of any considerable body of information seems to require

[2] "Our ineptitude in getting at the record is largely caused by the artificiality of systems...and information is found (when it is) by tracing it down from subclass to subclass."

that the information be pre-arranged in a classification system.

The classification of knowledge in the broad philosophical sense is as old as self-conscious knowledge itself; but the classification of books, items of literature, or items of information is a product of the nineteenth century. It is customary to explain the adoption of classification systems by libraries in terms of the growth of the open-shelf system of public and college libraries which occurred during the nineteenth century. If books are to be displayed for patrons to do their own browsing and make their own selection, the books must be arranged in some generalized subject order, e.g., science, religion, fiction, history, hobbies, etc.

This explanation does not illuminate the jump from generalized shelf arrangement to universal systems involving close and detailed classification of each book or of each item of information. Some other explanation is needed to account for the tremendous emphasis upon classification systems at the turn of the nineteenth and twentieth centuries. In a few generations, we have witnessed the development of the Dewey system, the Library of Congress system, the Cutter expansive system, the Universal Decimal Classification, the Brown Classification, the Colon Classification, the Bliss Classification, the U. S. Patent Office Classification, and a host of others which were born and have died in some local library or information center.

The explanation for this phenomenon is to be found in the temper of the times. The nineteenth century was the age of biology in the sense in which the seventeenth century was the age of physical science and the eighteenth century was the age of reason and enlightenment. This is not to say that there was not important work in the physical sciences in the nineteenth century. Maxwell, Faraday, Gibbs, Peano, Frege, Gauss, Helmholtz and dozens of other important chemists, mathematicians, and physicists lived and worked during this period, and there were important discoveries in all fields of science. But the ideas which were generalized beyond the laboratory and which established the intellectual climate of the age came from the science of biology. From biological analogs came the ideas of social evolution, the class struggle, survival of the fittest, the white man's burden, manifest destiny, and the iron law of wages. Since biology is a taxonomic science, a science of classes, it is reasonable to expect that librarians and other systematizers should employ the biological notions of taxonomy and hierarchical classes to organize their books and items of information.

These builders of classification systems might have built better systems had they the wit or wisdom to perceive that their own structures were products of an evolution which would in turn destroy them. Charles Sanders Pierce, who was a mathematician and one of the great creators of modern logic, and not a biologist, warned the system builders of his day that the better they built for the day, the shorter would be the life of what they built.[3] And Nietzsche, who took evolution seriously, recognized that all the values of his day must be superseded; that the superman must follow man in the chain of evolution. This insight, and perhaps a few biological complications, drove him mad.

For it is certainly the cream of the jest that this age of biology was also in fact the age of Victorian smugness. Evolution explained all development and all error up to the status quo, and some way, some how, evolution was supposed to cease in its highest product, nineteenth century Europe.

In 1937, writing on this same topic in collaboration with Dr. John Lund, we summed up the situation in these words:

> "The nineteenth century had an abiding faith in the permanence of its values and the ultimate validity of its scientific structures. This is illustrated by the belief of systematizers that, once a good classification of knowledge was achieved, it would be permanent. They did not learn from the fate of previous systems that their own must of necessity become obsolete.[4]
>
> > 'Decimal classification was born in a period when mankind had full confidence in the all-mightiness of materialistic wisdom. The middle of the nineteenth century was the culmination point of scientific positivism. It seemed that the totality of available knowledge as well of future knowledge could be arranged in a simple predetermined plan. Forgotten was the word of wisdom that Hamlet to Horatio spoke.....' "[5]

[3] Collected papers of Charles Sanders Pierce, p.83, edited by Charles Hartshorne and Paul Weiss, Harvard University Press, Cambridge.

[4] The Library Quarterly, 7:380 (1937).

[5] Dr. Donker Duyvis, in an address given before the British Society of Bibliography. Reported in "Notes and news," p.243.

Lest it be supposed that we were unfair then and unfair now in our estimate of the temper of the times, we quote below from the Annual Register of the University of Chicago for 1902. In the announcement of courses for the Department of Physics for that year, students electing physics were told:

> "While it is never safe to affirm that the future of Physical Science has no marvels in store even more astonishing than those of the past, it seems probable that most of the grand underlying principles have been firmly established, and that further advances are to be sought chiefly in the rigorous application of these principles to all the phenomena which come under our notice.
>
> It is here that the science of measurement shows its importance where quantitative results are more to be desired than qualitative work. An eminent physicist has remarked that the future truths of Physical Science are to be looked for in the sixth place of decimals."[6]

The chairman of the Department of Physics at that time was Professor A. A. Michelson of the famed Michelson-Morely experiment whose implications already threatened the stability of the "grand underlying principles [so] firmly established."

Is it any wonder, then, that "mere" librarians should delude themselves into thinking that they could classify all knowledge for all time? Dewey's proud boast that his classification system could include any new developments in knowledge through the device of adding another digit after the decimal point, is exactly on a par with the assertion that new developments in physical science would consist of refinements of measurement.

[6] Annual Register of the University of Chicago, p. 292 (1901-1902).

A few years ago, it could be said with considerable assurance that classification was a dead issue so far as librarianship and documentation were concerned. More and more librarians and scientists had come to depend on alphabetical subject-heading systems and alphabetical indexes. The excessive preoccupation of the FID with the Universal Decimal Classification was largely responsible for its lack of influence and effectiveness in practical librarianship and documentation.[7] Libraries already committed to various classification systems had come to regard such systems as devices for shelf notation and not as usable and viable keys to the subject content of their collections.

A few names, Bliss and Ranganathan; a few libraries, the John Crerar Library in Chicago and the Engineering Societies Library in New York retained throughout the first half of the present century a practical interest in problems of classification. But the general feeling on matters of classification has been well summed up by Dean Jesse Shera of Western Reserve University School of Library Science in the following passages:

> "Today, under the impact of a rapidly growing volume of graphic records, and the appearance of new forms of publication, traditional library classifications are becoming hopelessly inadequate. No amount of basic revision or tampering with their organic structure can save them from this failure. As guides to the

[7] Bibliographical Services: Their Present State and Possibilities of Improvement. Appendix, p. 12 (1950). The UNESCO Library of Congress Bibliographical Survey.

subject content of the library they are essentially meaningless. Even librarians, who are best qualified to interpret them and to exploit their virtues, use the notation only as a guide to location, and largely ignore the interdisciplinary relationships that they were designed to reveal. Yet, as their efficiency has declined, the cost of their maintenance has increased until at least one major research library has abandoned subject classification of its book stocks and has turned to other and more promising forms of bibliographic organization."[8]

"The history of library classification, then, has been the narrative of a pursuit of impossible goals, and its pages are strewn with the wreckage of those who either were blissfully unaware of the dangers by which their paths were beset, or who hoped to circumvent them through mere modification of previous schematisms or simple tinkering with notation. Today the essential failure of traditional library classifications is no more real than it was three-quarters of a century ago, but it has become more apparent because of the increasing bulk and complexity of the materials that libraries are being called upon to service, and the growing specialization of the demands that librarians are being asked to meet."[9]

8
"Classification as the Basic of Bibliographic Organization" in Bibliographic Organization, Papers Presented before the Fifteenth Annual Conference of the Graduate Library School, July 24-29, 1950, p. 72. Edited by Jesse H. Shera and Margaret E. Egan, Chicago, The University of Chicago Press, (1951).

9
"Classification: Current Functions and Applications to be the Subject Analysis of Library Materials," in The Subject Analysis of Library Materials, p. 32. Edited by Maurice F. Tauber, School of Library Service, Columbia University, New York (1953).

How then do we account for the renewed interest in classification as a method of information control? Within the last few years, we have witnessed the birth (and in some cases, the rapid death) of dozens of new classification systems, among which we can name, The Story Classification for the Army Technical Reference Service; the Office of Naval Research Project Status Classification; the Research and Development Board Classification of research projects, the American Society for Metals - Special Libraries Association Metallurgical Literature Classification, and the Standard Aeronautical Indexing System.[10] There has been a revival of interest in the Universal Decimal Classification, in the Patent Office Classification, and in Ranganthan's Colon Classification. A research project supported by Federal funds has labored for several years and is still laboring on the development of "abstraction ladders" and "semantic factoring".

This renewed search for the solution to an unsolvable problem results from a paradox, namely, the promise of machine organization and retrieval of information, and the actual slowness of the machine in the linear searching of an index. As we shall see in the following discussion, classification becomes one of the methods proposed for dividing an index in order to shorten the time required for a machine search.

10 In spite of the name, the Standard Aeronautical Indexing System is a hierarchical classification system.

PART II

Let us suppose we are searching for the name "Baker, Able Charlie" in a village telephone book containing about 1000 names. To search for this name might take a minute or two, occupied with picking up the book, finding the proper page and column, and scanning the proper column for the name being sought. Now it is quite practical to utilize an IBM machine, or some similar machine, or even a deck of edge-notched cards, to find one name in a random file of a thousand names, in about the same time required for the manual search of an alphabetical file in a minute or two. But suppose we are looking for the name "Baker, Able Charlie" in a list of a million names comparable to the New York telephone book. It might take us a little longer to lift the heavier book, to find the right page and the right column, and to scan by the given names and address as well as the last name. Nevertheless, the time required for a search for one name in a alphabetical list of a million names is of the same order of magnitude as the time required to find one name in an alphabetical list of a thousand names. But a machine search for one name in a random list of a million names will take one thousand times as long as a machine search for one name in a thousand.

It was the more or less vague realization of this fact that led the early advocates of the application of punched-card machines for the organisation and the retrieval of information to recognise that machine methods could not be applied efficiently to the random searching of large masses of information. No machine search of a large random list can approach the speed with which the mind can jump to the exact position in an ordered list. It would be silly to randomise a list of names in a phone book, or subject headings in an alphabetical index, in order to search for any particular name or heading with punched-card machines. An ordered list when it is over a certain size always enables the mind which recognises and utilizes the order to beat the machine. The conclusion to be drawn here is that contrary to popular misconceptions, the larger the number of qualitatively different units in a linear system of information, the less applicable are standard punched-card systems or even magnetic tape systems to the problem of searching; and this conclusion leads, in turn to a search for 1) ways to cut down the size of indexes and 2) ways to prefile or classify items of information.

The extent to which coordinate indexing cuts down the size of the index by eliminating the need for the alphabetizing of all permutations of terms in the indexing system, seems to offer the promise of efficient use of machine methods. Consider for example, a collection of 250,000 items to be organized in a system of information storage and retrieval. The items might be anything - documents, reports, patents, film footage or items of hardware in a supply catalog. With standard indexing systems, the size of the index would be 250,000 times the number of ways each item was indexed. Let us assume that an average of four terms is required to properly indexing or identify each item. The permutation of 4 is 24, and this gives us a maximum figure of 6,000,000,000 headings in the index. No index ever attempts to use all possible permutations of its terms as headings; but a barely adequate index in which each item is indexed by four terms will have at least four times the number of index headings as items. In this case, there would be 1,000,000 indexing headings in order to insure that each term used in the index will be in a filing position and will be found in proper alphabetical order. A punched-card system which could utilize one card per document and enter all four indexing terms of a document on this one card would require only 250,000 cards for an adequate index. The same reasoning can be applied to edge-notched card systems. Hence, it is true that the use of machine methods can reduce an index of 250,000 items from 250,000 x n (where n is the number of index entires per item) to 250,000 by eliminating the usual requirement of preparing multiple entries for each item. The elimination of multiple entries follows from the capacity of machine systems to search for an item under any word which indexes it and to combine all such words for an item on one card. We can see, therefore, that for systems in which n is large and the number of items is small, punched-card machines and edge-notched cards do offer the promise of a considerable reduction in the size of such systems. But note that the size in any instance cannot be smaller than the number of items in the system. This means that when a number of units is large, e.g., patent files, items of supply, intelligence reports, scientific and technical reports, case records in a large hospital, etc., ordinary systems of machine organization and retrieval are not practical. The random search of 1,000,000 items by standard punched-card systems will take about 33 hours per single search.[11] Surely the time required to find an item in the standard alphabetical index of a million items or even 10 million is only a fraction of this time. Of course, an expenditure of time is required to set up and maintain alphabetical indexes, but even if this time is more than that required to punch a set of cards to be maintained in random order, if any appreciable use is made of a random system, the great excess of time for machine searching will soon more than dissipate the savings, if any, realized in setting up the system. Finally, although some systems claim the possibility of asking multiple questions

[11] This figure is based on the ability of certain experimental IBM equipment to scan an entire card at a rate of 500 cards per minute. Standard, commercially available IBM equipment which sorts and selects a column at a time would require 33 hours for the first search by the first column. Selection by the second, third column, etc. would require an additional time, determined by the number of cards eliminated at each step of the search.

during a single search, the total system is searched for any question or questions and is unavailable for consultation by any other searcher. The conclusion to be drawn from the considerations outlined above is that the reduction in the size of a file made possible by coordinate indexing does not in itself establish the practicability of punched-card searching of large systems. Some more drastic reduction in size is required.

If instead of searching the file we collate or coordinate terms, we can enormously reduce the time of searching, but we will have to pay for this reduction in searching time by a compensatory increase in the size of the file. Consider the following arrays in which the letters represent ideas or terms in the index and the numbers represent the items to be indexed. Let us assume, as we did above, that each item is indexed by four terms.

Searching

1	A M N O
2	B C D T
3	A B M R
4	L N O P
5	C G H K
6	F G M P
7	L P R T
8	H K L S
9	B C R S
etc.	

Collating

A	1		3						
B		2	3						9
C		2			5				
D		2							
F						6			
G					5	6		8	
H					5				
K					5			8	
L				4			7	8	
M	1		3			6			
N	1			4					
O	1			4					
P				4		6	7		
R			3				7		9
S								8	9
T		2					7		
etc.									

In the array labelled <u>Searching</u>, the nine items and thirty-six indexing terms can be recorded on nine cards. A search for any item indexed under the term "G" , or any combination of terms "LP" , necessitates a search of the total file, in this case nine cards. But, if the array were continued on through a million items the search for "O" or "LP" would still involve the examination of the total file of 1,000,000. If we turn to the array labelled Collation, we note first that each of our nine numbers is repeated four times, a total of 36. This means that 36 punched cards (the number of items x n) rather than nine are needed. Now if we desire all items indexed with the term "O", no searching is required; the array shows us that item 5 and 6 are under "O". If we wish all items indexed under LP, we are not required to search the whole file, but only to compare (collate the numbers recorded) under "L and P", e.g., L 4-7-8- P 4-6-7-. It is apparent that items 4 and 7 are indexed under " LP".

In one respect, these arrays are misleading because they seem to indicate that there are more terms than documents. For any large system of information the reverse is true; there are always less terms than items in any system of information large enough to require any organization at all. Even in this small sample, the number of items under any term is less than the total number of items.

The number of cards required for setting up a collating system for 250,000 items is again equal to 250,000 x n (where n is the average number of terms used to index any item). Where n=4, we have a million cards just as we did in our standard alphabetical file. In collation, however, the cards are not maintained in a single array; instead we will have as many arrays as there are different terms used in the system. Thus, if 10,000 different terms are used in indexing 250,000 documents we will have 10,000 arrays of cards. Each array will then contain $\frac{250,000}{10,000} \times 4$ cards or 100. The total number of cards will be the number of arrays times the number of cards in each array: 10,000 x 100 = 1,000,000.

The process of collating for any item or items indexed by any four terms in this system will involve the collation of 400 cards (four arrays of 100 each). Since, in a file organized for selection by searching, we will have to search 250,000 cards, we can conclude that whereas we multiplied our file by 4, in order to shift from searching to collating, we cut machine time by a factor of $\frac{250,000}{400}$ or 625. On the other hand, a punched-card file for collating must be maintained in a fixed order. It is not possible to collate two random files by standard machine methods.

The superiority of collation to searching as a machine method for making selections from large systems does not materially advance us. A file used for collation equals in size a standard index and must exhibit the same type of rigorous order. Furthermore, collating is a relatively slow machine process, and collating four arrays involves three machine runs. There is no evidence that such an operation is less time-consuming and more efficient then searching for the proper heading by mind, eye, and hand in a regular index.

We are, of course, assuming that an item indexed by four items will require and receive only <u>four</u> entries in a standard index. If, in the standard index, we wish to provide for the other possible permutations of terms in an alphabetical sequence, we must increase the size of the file. In a collation system, we provide for all possible permutations with a maximum number of cards equal to the product of the number of items and the average number of terms used to index each item. Using our figures of 250,000 items and 4 terms, the difference here can be expressed by means of the following equations:

Collation: 250,000 x 4 = All possible permutations.

Standard Alphabetical Systems: 250,000 x 24 = All possible permutations.

This capacity to provide for all permutations without increasing the size of the file constitutes a definite advantage of machine systems over standard alphabetical indexes. But this advantage is only significant when the number of desired permutations is large and the number of units indexed is small.

From these considerations, the conclusion has been generally drawn that the linear machine searching of an alphabetical index is not a practical alternative to established manual methods. It is to escape this conclusion that once again attempts are being made to develop classification systems. A classification system which provides a hierarchical set of classes and subclasses presumably makes it possible to search for any item in a class or subclass, rather than in a total system. Consider, for example, the Patent Office Classification system which contains upwards of 2,500,000 patents in over 300 main classes and 44,000 subclasses. Suppose that each patent could be uniquely classified and that a search for a group of analogous patents could be restricted to a search of a single class. We could then set up 300 separate files of punched cards, and, in each of these files make a complete search in an average time of 16 minutes, whereas it would take over 80 hours to search all the cards. If the cards were prefiled in 40,000 subclasses, it would take less than a minute to find any single patent or group of patents in the same subclass.

Unfortunately, as is generally recognized, the Patent Office Classification does not accomplish the unique classification of analogous patents. Recently we made a test search for patents on "aircraft de-icing equipment." Even though we found a subclass "aircraft-ice removing equipment" under the main class "aircraft," much of the analogous art was found through cross references to be in the other classes, namely: "heat engines", "pumps", "vibrators", etc.

The fact that any search in the Patent Office Classification may involve a dozen or more classes and subclasses, indicates that this classification will not provide the unique location and mutually exclusive classes required for cutting down the time of machine searching. This conclusion is to be expected since the Patent Office Classification was not developed for use with machines.

Therefore, we must turn our attention to attempts to devise classification systems designed especially for machine searching. In this connection we will consider briefly Dr. Story's "Proposed Classification List for the Army Technical Reference Service"[12] and Mr. James Perry's work with "abstraction ladders" and "semantic factoring".

Dr. Story attempted to construct a classification which would permit the assignment of one class number to a report,[13] but in his own test he assigned an average of 1.935 class numbers per report,[14] thus admitting his attempt at creating mutually exclusive class numbers was unsuccessful. Further-

12 The material which follows is based on an administrative report by C. D. Gull to the Armed Services Technical Information Agency.

13 His rule for classifying is quote on p. 24-25 of a report, "Analysis of the Proposed Classification List for the Army Technical Reference Service, August, 1949", prepared by Documentation Incorporated in December 1952, for the Armed Services Technical Information Agency, Washington 25, D. C.

14 Op. cit., p. 32.

more, he recognized that some subjects are common to more than one main class, and provided three tables of numbers for those subjects,[15] many of which are not duplicated in the main classes. All of the numbers were recorded on IBM punched cards, and the cards arranged by main classes, which threw the numbers from the tables out of order. As a result, a search for any number from the three tables required a search of all the punched cards, and thus the attempt at reducing the number of cards to be searched by using a classification was vitiated for the tables as well as for the main classes.

Published reports are not available from which we can evaluate fully the work of Mr. Perry and his colleagues. Enough has been said publicly to permit us to offer a preliminary estimate of the contribution of the concepts of "abstraction ladders" and "semantic factoring", to a solution to this problem.

An "abstraction ladder" as originally proposed by Mr. Perry was a highly descriptive name for the taxonomic relationship of classes:

3	Phylum	Chordata
32	Class	Mammalia
327	Order	Carnivora
3274	Family	Canidae
32748	Genus	Dog
327485	Species	Labrador Retriever
327482	Species	Airdale

[15] Op. cit., p. [illegible].

The numbers at the left of the table can be considered code designations for the various levels of abstraction. Thus, if we wish to find everything on Mammalia, we sort, or search, for everything coded "32". Such a search would at the same time deliver all information on Carnivora (327), Canidae (3274), Dogs (32748), etc. We can also start at the other end of the ladder and search for Labrador Retrievers by searching for everything coded 327486. Such a search would not give us any information pertaining to the higher steps of the ladder. In short, an abstraction ladder makes possible a generic search, i.e., a search for classes within classes without requiring us to specify in advance all the classes and subclasses contained in the class which is the object of the search. Furthermore, if we prefile the material by class, i.e., all "3's" together or all "32's" together, presumably it is only necessary to search such prefiled groups for any information or any item belonging to the same abstraction ladder. To be sure, this possibility follows from a predetermination of abstraction ladders and a coding system based upon such predetermination.

Suppose that we have another section of this file on the subject, "Manners" and "Customs" and that under this class we have the following abstraction ladder:

9	Manners and Customs
92	Rural Sports
928	Hunting
9286	Hunting Dogs
92864	Labrador Retrievers

It is immediately obvious that all material on Dogs will not be found by searching for 32748, nor will all material on Labrador Retrievers be found in one abstraction ladder or pre-filed group under Chordata, Mammalia, or Dogs.

It will be said at this point that the taxonomic relations, the relations of class inclusion and subordination, illustrated by the abstraction ladder from Chordata to Labrador Retrievers are more "true" or more "objective" than the relations exhibited by the ladder Manners and Customs to Labrador Retrievers. In a certain sense, we admit this is the case. The apparent objectivity of abstraction ladders in the fields of zoology and botany is a reflection of the fact that in these fields there are established taxonomies which have achieved general acceptance. We must still question whether or not outside these special fields, there is any sense in regarding one abstraction ladder as more "objective" or "true" than another. For example the abstraction ladder:

Manners and Customs

Rural Sports

Hunting

Hunting Dogs

Labrador Retrievers

is no more objective or true than the abstraction ladder:

Canada

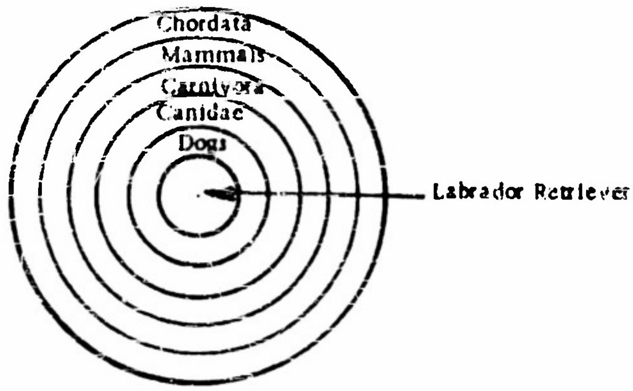
Newfoundland

Fauna

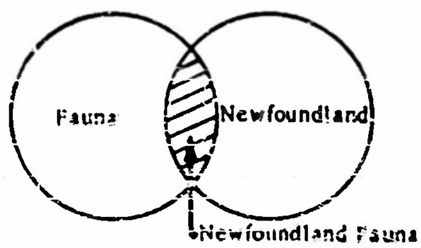
Labrador Retrievers

It will be said here that we have used different principles of subordination; that the sense in which Newfoundland is subordinate to Canada is not the sense in which Fauna is subordinated to Labrador; or the sense in which Rural Sports is subordinated to Manners and Customs. Again, we recognize that this vague sense of different kinds of subordination has a basis in fact, but what this basis is has not hitherto been specified in the literature of classification and information theory.

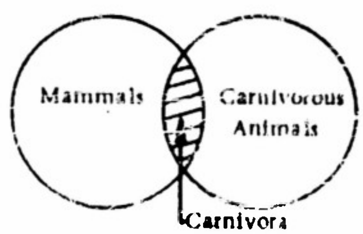
Subordination in classification systems is always a relation of class inclusion as contrasted with the relation of logical conjunction or coordination which is the characteristic relationship of elements in a coordinate index. An abstraction ladder from Chordata to Labrador Retrievers can be illustrated by a set of circles in which the subordinate circle is included in the higher or more generic circle.



On the other hand, the logical conjunction of Fauna and Newfoundland can be pictured as two overlapping circles:



In the ladder Chordata to Labrador Retriever, the class Mammals is included in the class Chordata, since there are no Mammals which are not Chordata; but there are carnivorous animals which are not Mammals. For example, birds, fish, and some reptiles. The Carnivora, as an order of the class Mammals, is a class formed by the logical conjunction of the class carnivorous animals with the class Mammals.



If two ideas or classes are related by the relation of logical conjunction we can set up the order of subordination in either direction as suits our purpose. We can for example, set up a ladder in which the largest class is carnivorous animals and we can subordinate Mammals to carnivorous animals, just as we can subordinate Newfoundland to Fauna; but Carnivora, defined as special class of Mammals having certain special characteristics is subordinate to Mammals by definition. There is no question of objectivity or truth here but only a matter of definition or the meanings of words. Being a mammal is part of the definition of being a Carnivora. This is more obvious, perhaps, in the case of Newfoundland Fauna, which by the very meaning of the words is recognized as a subclass of the class Fauna. The certainty and assurance with which we subordinate Carnivora to Mammals or Newfoundland Fauna to Fauna derives from our acceptance of formal definitions and not from any recognition or discovery of truth or true relationships in nature.

A taxonomy is a system of definitions which fixes the relationship between a set of terms and, hence, of a set of classes denoted by the terms. Systematic zoology and botany and parts of chemistry are taxonomic sciences because their vocabularies are fixed by such definitions. Now the extension of the taxonomic method to science in general, or to the field of intelligence, assumes that the vocabularies of science or intelligence constitute a taxonomy or a system of abstraction ladders. This assumption we can state without qualification is false.

PART III

In developing the theoretical discussion presented above, we examined several major classification systems to determine how the subordination of classes is achieved in them. The National Advisory Committee for Aeronautics classification system which was studied earlier in connection with the preparation of one of our reports to ASTIA, is not a suitable system for our present purposes since it is admittedly based on local and arbitrary ground rules. We have, therefore, taken our examples from systems which claim to be based on rational and objective considerations, namely the U. S. Patent Office Classification and Bliss's "Bibliographic Classification".[16] We might have chosen examples from the Library of Congress system or the Dewey system both of which we have subjected to thorough examination; but the Patent Office Classification seemed particularly appropriate to our purposes since those who are responsible for creating and using it make a great point of the inability of indexes to display the generic relationships and associations required in patent searching. Bliss's "Bibliographical Classification" has been chosen because it is the latest and, presumably, the most "scientific" of all library classification systems.

In both systems we discovered and distinguished three methods of achieving subordination of one idea to another:

[16] A Bibliographic Classification, by Henry Evelyn Bliss, H. W. Wilson Co., New York (1952).

the semantic, the topical, and the taxonomic. We also discovered that taxonomic subordination (or true classification) can only be found in the taxonomic sciences included as sections of over-all classification systems; and that the balance of a general classification system like the Patent Office system, or the Bliss system exhibited only semantic and logical subordination.

Semantic Subordination:

As the name indicates, semantic subordination is purely verbal in character and differs from alphabetical indexing only in being arranged differently on a page. Consider for example, the following sets of terms and phrases which might be found in any alphabetical index:

Functions, Additive, of aggregates
Functions, Continuous
Functions, Differentiable
Functions, Discontinuous
Functions, Integrable
Functions, Symmetric
Functions, Types of

or

Science
Science, History of
Science, Philosophy of

Science, Principles and methods of

or

Valves

Valves, Check

Valves, Gate

Valves, Reducing

Valves, Seated

If we arrange these sets of terms to look like parts of a classification system by utilising indentation on a page, as Mr. Bliss has done, we get the following:

Types of functions:

Aggregates of additive functions

Continuous functions

Differentiable functions

Discontinuous functions

Integrable functions

Symmetric functions

or

Science

History of science

Philosophy of science

Principles and methods of science

or

Valves

Check valves

Gate valves

Reducing valves

Seated valves

The following example of semantic subordination is taken from the Patent Office Classification, Class 192, "Clutches and Power-Stop Control." Under this classification, there is listed:

30 Clutches

30.5 Impact delivery type

31 Automatic

32 Manual control

41 One way engaging

48 Multiple

These headings can be rearranged for an alphabetical index as follows:

Clutches

Clutches - Impact delivery type

Clutches, Automatic

Clutches, Automatic - Manual control

Clutches, Automatic - One way engaging

Clutches, Multiple

Since the beginning of modern librarianship, exponents of classification have been able to convince a great many people that the indented arrangement is more logical than the inverted, whereas examination discloses only a difference in semantic or physical arrangement. Mr. Bliss and those responsible for the Patent Office Classification share a failure to recognize that classification, to the extent that it achieves subordination by semantic means (e.g., subordinates "check valves" to "valves", "discontinuous functions" to "functions" or automatic clutches to clutches, depends upon words and not upon any logic of ideas which underlies the words. That is to say, the boast which classifiers make of having achieved logical order as opposed to verbal or alphabetical order is empty and meaningless, to the extent that they use semantic subordination.

<u>Topical Subordination</u>:

The second way classifiers achieve subordination is through topical subdivision. This method is called "cross classification" by Mr. Bliss in his introduction and he illustrates it by means of the following tables:

	Plants	Insects	Birds	Mammals
Aquatic				
Terrestrial				
Amphibious				
Xeric				

	Aquatic	Land	Amphibious	Xeric
Insects				
Birds				
Plants				
Mammals				

It should be apparent that there is no real difference between these two tables and that it is no more logical or scientific to subdivide forms of life by habitat than to subdivide habitat by forms of life. Mr. Bliss realizes this; hence, his use of the term "cross classification" and his statement that: "Classes, or sub-classes, of the same grade, or order, of division are termed coordinate, and the principle of placing them in such order is coordination. Subordination and coordination are thus relative to division and gradation. The coordinate sub-classes of several coordinate classes may be coordinated".[17] However, he does not take the final and necessary step which is the recognition that the subordination of one topic to another is arbitrary and parochial and has no claim to logical or universal significance.

The following example of topical subordination is taken from the Patent Office Classification Class 75, "Metallurgy."

17 Bliss, Ibid., p. 6.

122	Alloys
138	Aluminum
139	Copper
140	Tin
141	Zinc
142	Magnesium
143	Silicon
144	Nickel
145	Silver
146	Zinc
147	Magnesium
148	Silicon
153	Copper
154	Tin
156	Lead
156.5	Zinc
157	Zinc
157.5	Zinc

Here again, it is clear that topical subordination is really coordination. There is no sense in which aluminum is more generic than copper or copper more generic than tin or zinc. We can put this same observation in stronger language by noting that it is nonsense to suppose an arrangement on a page can make copper generic to tin or tin subordinate to, or a subdivision of copper in the sense that carnivora are subordinate to or a subdivision of ma mals or iodine is subordinate to or a subdivision of halides.

These two forms of relationship, the topical and semantic, constitute overwhelming proportion of most classification systems. Once this premise is established the conclusion follows that universal classification is no more significant than a pattern of printing on a page, and has no logic other than the logic of general discourse.

Taxonomic Subordination:

All general classification systems which include sections on botany, zoology and chemistry exhibit, as we noted in Part II, genuine taxonomic relations of one-way subordination and inclusion. In the Patent Office Classification we find instances in Class 260, "Chemistry, Carbon Compounds", e.g.:

241	Azine
250	Diazine
252	Pyrimidine

In this case as in our previous examples drawn from the field of zoology, the very meaning of the words, determines that Azine includes Diazines and Diazine include Pyrimidines.

Since Bliss's Bibliographic Classification utilizes biological taxonomies for his class "F" Botany, and his class "G" Zoology, there is no need to labor this point any further.

PART IV

Although fully developed abstraction ladders do not exist outside of the special taxonomic sciences, in all fields of science we do use words which are defined in terms of their relation to more generic words or ideas. The fact that in recent months we have heard little about abstraction ladders from Mr. Perry and more about an operation known as "semantic factoring" is no doubt attributable to his recognition that even though we cannot create truly significant abstraction ladders, we can usually consider any class in relation to a higher class. For example, a bomber may be defined as a type of airplane, and in some information systems it might be worth while to index any material on bombers under both headings, "bombers" and "airplanes". Such two-level relationships are not ladders except in a sense which is so minimal as to be trivial. We have noted in our empirical work that indexing on two levels usually makes a good deal of sense, whereas the attempt to go beyond two levels makes very little sense. In one instance, for example, we determined to index material on pentodes under the heading "tubes", but it would have seemed silly to use the next higher level "electronic devices". We might index material on iodine under "halogens" but not under "chemicals". In a lecture given at the School of Library

Service, Columbia University in 1953, Mr. Perry used as an example of semantic factoring, the relationship between the terms "weapon" and "mine". He could not offer any term for the next higher or lower level and stated that semantic factoring usually involved only two levels. One fact which may account for the ease with which we can index on two levels, or our readiness to accept semantic factoring as a two-level process, is the standard dictionary practice of defining something by describing it as a special kind of the next higher genus. Webster's Collegiate Dictionary defines "pistol" as a certain kind of firearm; it defines "firearm" as a certain kind of weapon; and it defines "weapon" as a certain kind of instrument.

But our problem here is not with the possibility of defining words on two or more levels, but with estimating the contribution of such definitions to the solution of the problem of excessive machine-search time. If in the vocabulary of any system of information the number of semantic factors is a small portion of the total vocabulary and each term in this vocabulary can only be factored one way, i.e., can be related to only one of the semantic factors, then we could appreciably reduce searching time by restricting searching for any item to an array of terms under one factor. But we have no reason to assume that this situation holds or could be made to hold for any actual information system. We recognise that pistols are firearms and firearms are weapons. But pistols may, with as much logic, be grouped with Sam Brown belts and clothing under the general class "officers' equipment" as with mines, atomic bombs and guided missiles under the general class "weapons".

In short, we may take advantage of what the dictionary tells us about the meaning of words to make our indexing more useful for searches at various levels of generality, but we cannot utilize such definitions to divide a system of information in order to reduce the time of search. Our conclusion here as with all of our conclusions in this paper, has an empirical as well as a theoretical basis.

In the paper we quoted at the beginning of this discussion, and in another paper in the same volume, we attempted to find in the idea of "categorization" a kind of special grouping of ideas or semantic factors without systematic subdivisions - a middle ground as it were between classification systems and the straight alphabetization of an index. We are now certain on the basis of efforts extending over a year and a half, and supporting evidence from one of our clients who also attempted to categorise an extensive vocabulary, that categorization is not practical for large systems. For particular and local purposes involving highly specialised collections of information, it is possible to establish relatively adequate categories, or semantic factors, or even abstraction ladders. But the more general the system and the more general its use, the more difficult it becomes to set up adequately defined and mutually exclusive categories.

If we were concerned with indexing all materials on a particular disease we might set up the following list of categories: symptoms, etiology, treatment, geographical distribution, prognosis, economic and social factors, complications, age groups, or any others which suggest themselves as generic interests or factors under which we could group our vocabulary of indexing terms. But we doubt the possibility of devising a set of categories or semantic factors for the terms used in indexing the subject of medicine, the field of science, the interests of the Department of Defense, or the claims of patents. This is not a counsel of despair, but rather realism in the face of experience.

In a certain sense, this paper may be regarded as a clearing of the underbrush before beginning construction. This metaphor is sound because it should warn us that underbrush is not something we can eliminate once and for all. We started in the beginning of this paper with certain quotations from Dr. Shera's work and certain observations which seemed to indicate that the choking underbrush of hierarchical classification systems had been recognized [illegible] and [illegible] to elimination; but then we [illegible] the [illegible] of [illegible] quirements of machine storage and [illegible] grew again in the form of "abstraction ladders", "[illegible] factoring", and "categorization".

If the mind is to recognize and [illegible] as [illegible] associations are formed in fact, it cannot be confined to a system of definitions which predetermine, [illegible], the "proper", "correct", or "true" association of facts or ideas. Classification systems and abstraction ladders are systems of predetermined associations which tend to [illegible] and falsify the full diversity and [illegible] of the mind's operations as it seeks meaning in the free association, dissociation, and reassociation of ideas.